((DICTIONARY creating maximum size) and (removing and stop and words)) and (removing and duplicate and words)
Terms used:
DICTIONARY creating maximum size removing stop words removing duplicate words

Found **16** of **60** searched out of **250,175**.

Sort results by: relevance

Display results: expanded form

Save results to a Binder

Open results in a new window

Refine these results with Advanced Search

Try this search in The ACM Guide

Results 1 - 16 of 16

1 Adaptive web-page content identification

John Gibson, Ben Wellner, Susan Lubar

November 2007 WIDM '07: Proceedings of the 9th annual ACM international workshop on Web information and data management

Publisher: ACM

Full text available: pdf(488.02 KB)

Additional Information: full citation, abstract, references, index terms

Bibliometrics: Downloads (6 Weeks): 21, Downloads (12 Months): 148, Citation Count: 0

Identifying which parts of a Web-page contain target content (e.g., the portion of an online news page that contains the actual article) is a significant problem that must be addressed for many Web-based applications. Most approaches to this problem ...

Keywords: conditional random fields, content identification, maximum entropy markov models, sequence labeling

2 Efficiently linking text documents with relevant structured information

Venkatesan T. Chakaravarthy, Himanshu Gupta, Prasan Roy, Mukesh Mohania
September 2006 VLDB '06: Proceedings of the 32nd international conference on Very large data bases
Publisher: VLDB Endowment
Full text available: pdf(598.27 KB) Additional Information: full citation, abstract, references, index terms

Bibliometrics: Downloads (6 Weeks): 20, Downloads (12 Months): 98, Citation Count: 0

Faced with growing knowledge management needs, enterprises are increasingly realizing the importance of interlinking critical business information distributed across structured and unstructured data sources. We present a novel system, called EROCS, for ...

3 Stemming Indonesian: A confix-stripping approach

Mirna Adriani, Jelita Asian, Bobby Nazief, S. M.M. Tahaghoghi, Hugh E. Williams
December 2007 ACM Transactions on Asian Language Information Processing (TALIP), Volume 6 Issue 4
Publisher: ACM
Full text available: pdf(369.46 KB) Additional Information: full citation, abstract, references, index terms

Bibliometrics: Downloads (6 Weeks): 34, Downloads (12 Months): 153, Citation Count: 0

Stemming words to (usually) remove suffixes has applications in text search, machine translation, document summarization, and text classification. For example, English stemming reduces the words "computer," "computing," "computation," and "computability" ...

Keywords: Indonesian, information retrieval, stemming

4 Characterization of national Web domains

Ricardo Baeza-Yates, Carlos Castillo, Efthimis N. Efthimiadis
May 2007 ACM Transactions on Internet Technology (TOIT), Volume 7 Issue 2
Publisher: ACM
Full text available: pdf(1.41 MB) Additional Information: full citation, abstract, references, index terms

Bibliometrics: Downloads (6 Weeks): 27, Downloads (12 Months): 273, Citation Count: 1

During the last few years, several studies on the characterization of the public Web space of various national domains have been published. The pages of a country are an interesting set for studying the characteristics of the Web because at the same ...

Keywords: Web characterization, Web measurement

5 Optimizing relevance and revenue in ad search: a query substitution approach

Filip Radlinski, Andrei Broder, Peter Ciccolo, Evgeniy Gabrilovich, Vanja Josifovski, Lance Riedel

July 2008 SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval

Publisher: ACM

Full text available: pdf(273.06 KB) Additional Information: full citation, abstract, references, index terms

Bibliometrics: Downloads (6 Weeks): 0, Downloads (12 Months): 0, Citation Count: 0

The primary business model behind Web search is based on textual advertising, where contextually relevant ads are displayed alongside search results. We address the problem of selecting these ads so that they are both relevant to the queries and profitable ...

Keywords: online advertising, relevance, revenue

6 Sources of Success for Boosted Wrapper Induction

David Kauchak, Joseph Smarr, Charles Elkan

December 2004 The Journal of Machine Learning Research, Volume 5

Publisher: MIT Press

Full text available: pdf(281.46 KB) Additional Information: full citation, abstract, references, cited by, index terms

Bibliometrics: Downloads (6 Weeks): 4, Downloads (12 Months): 33, Citation Count: 2

In this paper, we examine an important recent rule-based information extraction (IE) technique named Boosted Wrapper Induction (BWI) by conducting experiments on a wider variety of tasks than previously studied, including tasks using several collections ...

7 Robust and efficient fuzzy match for online data cleaning

Surajit Chaudhuri, Kris Ganjam, Venkatesh Ganti, Rajeev Motwani

June 2003 SIGMOD '03: Proceedings of the 2003 ACM SIGMOD international conference on Management of data

Publisher: ACM

Full text available: pdf(271.47 KB) Additional Information: full citation, abstract, references, cited by, index terms

Bibliometrics: Downloads (6 Weeks): 44, Downloads (12 Months): 296, Citation Count: 35

To ensure high data quality, data warehouses must validate and cleanse incoming data tuples from external sources. In many situations, clean tuples must match acceptable tuples in *reference tables*. For example, product name and description fields ...

8 Manageability, availability, and performance in porcupine: a highly scalable, cluster-based mail service

Yasushi Saito, Brian N. Bershad, Henry M. Levy
August 2000 ACM Transactions on Computer Systems (TOCS), Volume 18 Issue 3
Publisher: ACM
Full text available: pdf(2.52 MB) Additional Information: full citation, abstract, references, cited by, index terms

Bibliometrics: Downloads (6 Weeks): 22, Downloads (12 Months): 83, Citation Count: 2

This paper describes the motivation, design and performance of Porcupine, a scalable mail server. The goal of Porcupine is to provide a highly available and scalable electronic mail service using a large cluster of commodity PCs. We designed Porcupine ...

Keywords: cluster, distributed systems, email, group membership protocol, load balancing, replication

9 Communications of the ACM: Volume 51 Issue 2

February 2008 issue Volume 51 Issue 2
Publisher: ACM
Full text available: pdf(3.89 MB) digital edition Additional Information: full citation

Bibliometrics: Downloads (6 Weeks): 549, Downloads (12 Months): 1641, Citation Count: 0

10 Communications of the ACM: Volume 51 Issue 8

August 2008 issue Volume 51 Issue 8
Publisher: ACM
Full text available: pdf(6.85 MB) digital edition Additional Information: full citation

Bibliometrics: Downloads (6 Weeks): 0, Downloads (12 Months): 0, Citation Count: 0

11 RCV1: A New Benchmark Collection for Text Categorization Research

David D. Lewis, Yiming Yang, Tony G. Rose, Fan Li
December 2004 The Journal of Machine Learning Research, Volume 5
Publisher: MIT Press

Full text available: pdf(628.29 KB) Additional Information: full citation, abstract, references, cited by, index terms, review

Bibliometrics: Downloads (6 Weeks): 38, Downloads (12 Months): 223, Citation Count: 57

Reuters Corpus Volume I (RCV1) is an archive of over 800,000 manually categorized newswire stories recently made available by Reuters, Ltd. for research purposes. Use of this data for research on text categorization requires a detailed understanding ...

12 The string B-tree: a new data structure for string search in external memory and its applications

Paolo Ferragina, Roberto Grossi
March 1999 Journal of the ACM (JACM), Volume 46 Issue 2
Publisher: ACM

Full text available: pdf(363.37 KB) Additional Information: full citation, abstract, references, cited by, index terms

Bibliometrics: Downloads (6 Weeks): 58, Downloads (12 Months): 400, Citation Count: 30

We introduce a new text-indexing data structure, the String B-Tree, that can be seen as a link between some traditional external-memory and string-matching data structures. In a short phrase, it is a combination of B-trees and Patricia ...

Keywords: B-tree, Patricia trie, external-memory data structure, prefix and range search, string searching and sorting, suffix array, suffix tree, text index

13 Précis: from unstructured keywords as queries to structured databases as answers

Alkis Simitsis, Georgia Koutrika, Yannis Ioannidis
January 2008 The VLDB Journal — The International Journal on Very Large Data Bases, Volume 17 Issue 1
Publisher: Springer-Verlag New York, Inc.

Full text available: pdf(2.52 MB) Additional Information: full citation, abstract, index terms

Bibliometrics: Downloads (6 Weeks): 13, Downloads (12 Months): 51, Citation Count: 0

Précis queries represent a novel way of accessing data, which combines ideas and techniques from the fields of databases and information retrieval. They are free-form, keyword-based, queries on top of relational databases that generate entire multi-relation ...

Keywords: Free-from queries, Keyword search, Query processing

14 Shape-based retrieval and analysis of 3D models

Thomas Funkhouser, Michael Kazhdan
August 2004 SIGGRAPH '04: ACM SIGGRAPH 2004 Course Notes
Publisher: ACM
Full text available: pdf(12.56 MB) Additional Information: full citation, abstract

Bibliometrics: Downloads (6 Weeks): 111, Downloads (12 Months): 611, Citation Count: 0

Large repositories of 3D data are rapidly becoming available in several fields, including mechanical CAD, molecular biology, and computer graphics. As the number of 3D models grows, there is an increasing need for computer algorithms to help people find ...

15 The Web as a parallel corpus

Philip Resnik, Noah A. Smith
September 2003 Computational Linguistics, Volume 29 Issue 3
Publisher: MIT Press
Full text available: pdf(539.83 KB) Additional Information: full citation, abstract, references, cited by, index terms

Bibliometrics: Downloads (6 Weeks): 41, Downloads (12 Months): 258, Citation Count: 25

Parallel corpora have become an essential resource for work in multilingual natural language processing. In this article, we report on our work using the STRAND system for mining parallel text on the World Wide Web,first reviewing the original algorithm ...

16 The elements of nature: interactive and realistic techniques

Oliver Deusen, David S. Ebert, Ron Fedkiw, F. Kenton Musgrave, Przemyslaw Prusinkiewicz, Doug Roble, Jos Stam, Jerry Tessendorf
August 2004 SIGGRAPH '04: ACM SIGGRAPH 2004 Course Notes
Publisher: ACM

Full text available: pdf(17.65 MB) Additional Information: full citation, abstract, cited by

Bibliometrics: Downloads (6 Weeks): 268, Downloads (12 Months): 1447, Citation Count: 1

This updated course on simulating natural phenomena will cover the latest research and production techniques for simulating most of the elements of nature. The presenters will provide movie production, interactive simulation, and research perspectives ...

Results 1 - 16 of 16